ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Multivariate Analysis 94 (2005) 344–365

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Nonlinearity effects in multidimensional scaling

John C. Gower*, Roger F. Ngouenet[1]

Department of Statistics, The Open University, Walton Hall, Milton Keynes, England MK7 6AA, UK

Received 5 September 2003

Abstract

When multidimensional scaling of n cases is derived from dissimilarities that are functions of p basic continuous variables, the question arises of how to relate the values of the variables to the configuration of n points. We provide a methodology based on nonlinear biplots that expresses nonlinearity in two ways: (i) each variable is represented by a nonlinear trajectory and (ii) each trajectory is calibrated by an irregular scale. Methods for computing, calibrating and interpreting these trajectories are given and exemplified. Not only are the tools of immediate practical utility but the methodology established assists in a critical appraisal of the consequences of using nonlinear measures in a variety of multidimensional scaling methods.

AMS 1991 subject classification: 62-09; 62H99

Keywords: Biplots; Graphical representations; Matrix approximation; Multidimensional scaling; Nonlinearity; Prediction

1. Introduction

The following notation is used: $\{a_{ij}\}$ denotes a matrix $\mathbf{A}$ with typical element a_{ij}. The rows (columns) of $\mathbf{A}$, and vectors related to them, will be represented by row (column) vectors. $\mathbf{1}$ denotes a column-vector of ones, whose length is determined by context; $\mathbf{e}_k$ denotes a unit row-vector of length p, zero except for a unit kth element.

The multidimensional scaling (MDS) of a given $n \times n$ dissimilarity matrix with elements $\{d_{ij}\}$ gives a graphical representation $\mathbf{Z}$ of n points in an r-dimensional space, usually,

* Corresponding author.

E-mail addresses: j.c.gower@open.ac.uk (J.C. Gower), nroger2000@yahoo.com, roger@xlsolutions-corp.com (R.F. Ngouenet).

[1] Current address: XLSolutions Corporation, P.O. Box 9514, Seattle WA 98109.

0047-259X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2004.05.008

but not necessarily, Euclidean. The distance between two of these points approximates a function of d_{ij}. When the dissimilarities are derived from a data-matrix $\mathbf{X}$ of n rows, representing samples, and p columns, representing variables, the question arises of how to associate values of the variables with the MDS representation of the samples. When d_{ij} refer to Pythagorean distances (see below) linear biplots [3], suitably calibrated, may be used. The directions of linear biplots may be found as the columns of $\mathbf{B}$ that minimise $\|\mathbf{X}-\mathbf{ZB}\|^2$ (see e.g. [11] or [7]). This, so-called, regression method is often used, even when the d_{ij} are not Pythagorean and hence linearity assumptions are inappropriate or, at least, questionable. This paper provides a methodology for investigating the effects of nonlinearity in interpreting MDS analyses. We begin by reviewing the main ideas first presented by Gower and Harding [8], starting with what may be an unfamiliar perspective of classical linear biplots. Then, in Section 2, we show how the underlying notions may be generalized for nonlinear methods. A detailed account of these basic ideas, with extensions, is available in Gower and Hand [7]. Further extensions are given in Section 3. Finally, in Section 4, an example demonstrates how nonlinearity manifests itself in MDS.

The basic biplot derives from the Cartesian representation of the variables by p orthogonal coordinate axes and the samples by points in R^p. The position, relative to the axes, of the sample with values $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is given by the vector-sum $\sum_{k=1}^{p} x_{ik}\mathbf{e}_k$. The kth Cartesian axis is the locus of $\mu\mathbf{e}_k$ as μ varies and the value x_{ik} of the ith sample on the kth axis is given by the projection $\mathbf{x}_i\mathbf{e}_k'\mathbf{e}_k = x_{ik}\mathbf{e}_k$. In the usual way, R^p is endowed with a distance d_{ij} where $d_{ij}^2 = (\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)'$. For reasons which will become clear, we prefer to refer to d_{ij} defined in this way as Pythagorean distance, and reserve the term Euclidean distance for Euclidean embeddable distances.

In statistical work, methods are sought that approximate this exact p-dimensional representation by a more convenient low-dimensional space of r dimensions, where often $r = 2$. The classical way of obtaining this approximation is to consider an r-dimensional subspace R^r of R^p spanned by unit vectors given as the columns of an orthonormal matrix $\mathbf{V}_r$. The projections onto R^r of the points representing the samples are $\mathbf{Z} = \mathbf{XV}_r\mathbf{V}_r'$ and R^r is chosen to minimise $\|\mathbf{X}-\mathbf{Z}\|^2$ the sum-of-squares of the difference between the points representing the samples and their projections. [2] This is Principal Components Analysis and, as is well known, $\mathbf{V}_r$ is given by the r dominant eigenvectors of $\mathbf{X}'\mathbf{X}$ or, equivalently, the r dominant right singular vectors of the singular value decomposition $\mathbf{X} = \mathbf{U\Sigma V}'$. The representation $\mathbf{z}_i$ in R^r of the ith sample $\mathbf{x}_i$ has coordinates in R^p given by the projection $\mathbf{z}_i = \mathbf{x}_i\mathbf{V}_r\mathbf{V}_r' = \sum_{k=1}^{p} x_{ik}\mathbf{e}_k\mathbf{V}_r\mathbf{V}_r' = \sum_{k=1}^{p} x_{ik}\mathbf{b}_k$ where $\mathbf{b}_k = \mathbf{e}_k\mathbf{V}_r\mathbf{V}_r'$ are row-vectors, known as the biplot axes, that define p nonorthogonal directions in R^r. It may be verified that this setting of $\mathbf{b}_k$ agrees with that given by the regression method with $\mathbf{Z} = \mathbf{XV}_r\mathbf{V}_r'$. Just as one unit of the kth variable in $R^p$is represented by $\mathbf{e}_k$, in R^r one unit is represented by $\mathbf{b}_k$, even though this is not a unit vector. Note that, the vector sum $\sum_{k=1}^{p} x_k\mathbf{b}_k$ allows any sample, not necessarily one of the rows of $\mathbf{X}$, to be placed in R^r. This operation of placing a point in R^r is termed interpolation by Gower and Hand [7].

[2] Now $\mathbf{Z}$ has p columns, although it remains of rank r. Note that $\mathbf{XV}_r\mathbf{V}_r'$ and $\mathbf{XV}_r$ represent the same r-dimensional configuration, the first relative to the p original orthogonal axes in R^p and the second relative to r orthogonal axes in a subspace R^r spanned by the columns of $\mathbf{V}_r$.

The converse operation of associating a set of values with a given point in R^p, and particularly in its subspace R^r, is termed prediction and is given by projection onto the axes $\mathbf{e}_k$ $(k = 1, 2, \ldots, p)$. For any point $\mathbf{x}$ in R^r we have that $\mathbf{x} = \mathbf{x}\mathbf{V}_r\mathbf{V}_r'$. For the kth variable the value x_k is given by projection onto $\mathbf{e}_k$. Thus

$$x_k = \mathbf{x}\mathbf{e}_k' = \mathbf{x}\mathbf{V}_r\mathbf{V}_r'\mathbf{e}_k' = \mathbf{x}\mathbf{b}_k'.$$

For projection onto the kth biplot axis we have

$$\lambda_k \mathbf{x}\mathbf{b}_k'\mathbf{b}_k = \lambda x_k \mathbf{b}_k,$$

where the normalising factor $\lambda_k^{-1} = \mathbf{e}_k\mathbf{V}_r\mathbf{V}_r'\mathbf{e}_k'$. Thus, apart from the scaling λ_k, projection onto $\mathbf{b}_k$ in R^r is the same as projection onto $\mathbf{e}_k$ in R^p. The scaling factors are easily subsumed in labelling the points on the biplot axes, as described below. The crucial point is that the biplot axes allow interpolation and prediction using only information available within R^r.

Next we consider calibration of axes. Calibration is implicit in the definition of Cartesian axes with the vectors $\mathbf{e}_k$ giving the unit point for the kth variable. This aspect gets little emphasis in theoretical developments but is fundamental to scientific work and applied statistics. Of course, rather than calibrating the kth axis with markers labelling integer multiples of $\mathbf{e}_k$, it will often be more convenient to place markers at fractional or multiple quantities of $\mathbf{e}_k$, depending on the scale and range of the values of the kth variable actually occurring in the data under consideration. The result $\mathbf{z}_i = \sum_{k=1}^{p} x_{ik}\mathbf{b}_k$ shows that for interpolation, the point $\mathbf{b}_k$ may be labelled as a unit marker on the kth biplot axes and that $\mathbf{z}_i$ may be placed in R^r by vector summation, just as $\mathbf{x}_i$ can be placed in R^p. Similarly, we have seen that projection onto $\mathbf{b}_k$, apart from the scaling factor λ, gives correct predictions. It follows that, if the scaling factor is taken into account, the biplot axis may be calibrated to give correct predictions, again just as with the original Cartesian axes. In R^p we have that $\mathbf{V}_r$ is an orthogonal matrix, giving $\lambda_k = 1$, so that the calibrations for interpolation and prediction coincide, but in R^r they differ.

2. Non-Pythagorean distances

Although Pythagorean distances are often used explicitly or implicitly, as in principal components analysis, for data analysis, many other types of distance or dissimilarity are also used (see e.g. [9]). Then, many methods of MDS allow the samples to be exhibited by coordinates $\mathbf{Z}$ but the representation of the variables is less straightforward. The regression method may always be used to superimpose linear axes on the MDS but its performance with non-Pythagorean distances is questionable and needs examination. The methodology described below gives tools that help explore the effects of non-Pythagorean distances.

2.1. An exact representation for Euclidean embeddable distances

Methods for generalizing the linear Pythagorean case can be very general indeed but explicit algebraic results are available only when certain assumptions are made. We consider

more general considerations in the discussion of Section 5, but in the following we assume:

(i) *Additive squared distance*: Squared distance (written ddistance) between samples i and j is given by

$$d_{ij}^2 = \sum_{k=1}^{p} d_k(x_{ik}, x_{jk}), \tag{1}$$

where $d_k(x_{ik}, x_{jk})$ is a non-negative symmetric function applicable to the kth variable.

(ii) *Euclidean embeddability*. All $\binom{n}{2}$ d_{ij}^2 ddistances may be generated by the Euclidean distances between all pairs of n points whose coordinates are given as the rows of some matrix $\mathbf{Y}$.

The additivity assumption imposes some restrictions on the choice of distance but encompasses many distances in common use (see e.g. [9]). In the example of Section 4 we examine the case where

$$d_k(x_{ik}, x_{jk}) = \left(\frac{x_{ik} - x_{jk}}{x_{ik} + x_{jk}}\right)^2.$$

The embeddability assumption implies that the points $\mathbf{Y}$ lie in a Euclidean space R^m, $m \leqslant n-1$. The nonnegativity assumption for each $d_k(x_{ik}, x_{jk})$ combined with (ii) implies that the ddistances $d_k(.,.)$ for each variable separately are Euclidean. This follows from noting that when $d_{ij} = 0$, then the nonnegativity of $d_k(x_{ik}, x_{jk})$ in (1) implies that $d_k(x_{ik}, x_{jk}) = 0$ for $k = 1, 2, \ldots, p$. In particular, for Euclidean distances, $d_{ij} = 0$ when $x_{ik} = x_{jk}$, $k = 1, 2, \ldots, p$ and hence it follows that then $d_k(x_{ik}, x_{jk}) = 0$ for each variable. By considering two samples which are identical except for the kth variable, we have that $d_{ij}^2 = d_k(x_{ik}, x_{jk})$ so that (ii) implies that the ddistances $d_k(.,.)$ for each variable separately are Euclidean. Often, the function $d_k(.,.)$ will be independent of the choice of variable and then the suffix k may be dropped.

Writing $\mathbf{N}$ for the matrix $\mathbf{1}\mathbf{1}'/n$ and $\mathbf{D} = \{-\frac{1}{2}d_{ij}^2\}$ for the $n \times n$ ddistance matrix, Schoenberg [11] showed that

$$\mathbf{Y}\mathbf{Y}' = (\mathbf{I} - \mathbf{N})\mathbf{D}(\mathbf{I} - \mathbf{N}) = \mathbf{B}, \text{ say}, \tag{2}$$

is positive-semi-definite. Thus, $\mathbf{Y}$ is centred at the centroid of the n points and coordinate representations of its different orientations may be derived from any decomposition satisfying (2). In practice, the orientation given by the spectral decomposition $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$ is chosen, where $\mathbf{V}$ is the orthogonal matrix of eigenvectors and $\mathbf{\Lambda}$ is diagonal and gives the eigenvalues, assumed presented in non-ascending order. Then $\mathbf{1}'\mathbf{Y} = 0$, $\mathbf{Y} = \mathbf{V}\mathbf{\Lambda}^{1/2}$ and $\mathbf{Y}'\mathbf{Y} = \mathbf{\Lambda}$. With this centring and orientation, the points given by the rows of $\mathbf{Y}$ are expressed as principal coordinates [4,14] i.e. they are referred to principal axes through the centroid of the n points.

The original Cartesian axes, for which the rows of $\mathbf{X}$ are coordinates, have a representation in R^m in which the kth axis may be represented as a trajectory ξ_k ($k = 1, 2, \ldots, p$) defined as the locus as μ varies of the mapping in R^m of the pseudo-sample $\mu\mathbf{e}_k$. Usually, $m = n-1$ but it may be smaller as, for example, with Pythagorean distance where $m = p$. Gower and Harding [8] showed that the column-vector of ddistances of the mapped-pseudo sample

from the points given by $\mathbf{Y}$ are

$$\sum_{h=1}^{p} \mathbf{c}_h - \mathbf{c}_k + \mathbf{d}_k, \tag{3}$$

where $\mathbf{c}_h = \{d_h(0, x_{ih})\}$, $\mathbf{d}_k = \{d_k(\mu, x_{ik})\}$ and then, the coordinates of the mapped pseudo-sample are

$$\left.\begin{aligned} \mathbf{y}_\mu &= \tfrac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{Y}'\left(\sum_{h=1}^{p} \mathbf{c}_h - \mathbf{c}_k + \mathbf{d}_k - \tfrac{2}{n}\mathbf{D1}\right) \\ y^2_{m+1}(\mu) &= \tfrac{1}{n}\mathbf{1}'\mathbf{d}_k + \tfrac{1}{n^2}\mathbf{1}'\mathbf{D1} - \mathbf{y}'_\mu\mathbf{y}_\mu \end{aligned}\right\}. \tag{4}$$

The coordinate $y_{m+1}(\mu)$ in an extra $(m+1)$th dimension, needed to accommodate the new point, is potentially an embarrassment, because this extra dimension is needed for every value of μ, leading to an infinite-dimensional space. Arguments given by Gower and Hand [7, p. 262] show that, for our purposes, we may proceed as if there were only one extra dimension; this enlargement of R^m we label R^m_+. Thus, as μ varies, (4) is the parametric equation of the kth trajectory ξ_k in R^m_+ $(k = 1, 2, \ldots, p)$.

We calibrate ξ_k with markers at intervals given by $\mu = 0, \pm1, \pm2, \pm3, \ldots$ (or whatever other intervals are convenient) to indicate the values of the kth variable x_k; in the usual way for coordinate axes, visual interpolation between markers is required for intermediate quantities. The only term in (3) which depends on μ is the vector $\mathbf{d}_k$ and therefore the point on the trajectory that is nearest the point which represents the ith sample occurs when $d_k(\mu, x_{ik})$ is minimum. Because each $d_k(.,.)$ is separately Euclidean, as was shown above, the minimum is zero and occurs when $\mu = x_{ik}$. It follows that the values of the variables for the ith sample represented by a point P_i, say, are given by the nearest markers on each of the k axes, and that for continuous trajectories this is given by dropping the normals from P_i onto the trajectories ξ_k $(k = 1, 2, \ldots, p)$; this is termed normal projection. Thus, the values of the variables associated with a sample point are obtained by reading off the nearest markers on each axis; this process may be extended to all points of R^m_+. A further consequence of the additivity assumption (i) is that the position of a point is the vector sum of the p corresponding markers [8]. In this way, the trajectories ξ_k $(k = 1, 2, \ldots, p)$ are acting like conventional Cartesian coordinate axes both by using vector sums for placing samples in R^m_+ (interpolation) and by using normal projection for assigning values of the variables to any point in R^m_+ (prediction).

2.2. Approximation

Section 2.1 was concerned solely with the exact representations of the samples and nonlinear trajectories in R^m_+. Biplots are concerned with approximating this geometry of R^m_+ in an r-dimensional subspace L. L is a space that holds n points $\mathbf{Z}$, without loss of generality centred at their centroid, produced by some form of MDS, so that the distances generated by $\mathbf{Z}$ approximate the d_{ij} generated by $\mathbf{X}$, and equally $\mathbf{Y}$. When the MDS method is classical scaling/principal coordinates analysis, then $\mathbf{Z}$ is the orthogonal projection of $\mathbf{Y}$ given by its first r columns $\mathbf{Y}_r$. Then L may be identified with R^r of Section 1 and might

have been written R_+^r. With other MDS methods, L has to be embedded in R_+^m (see [10] for examples of how this is done); thereafter the MDS method used is immaterial.

Biplot axes are trajectories β_k in L that approximate ξ_k in R_+^m and therefore give information on $\mathbf{X}$ and the values of the observed variables. As with the linear case, these are of two kinds. Interpolative biplots allow new samples to be interpolated into L and for principal coordinates were shown by Gower and Harding [8] to be given by projecting ξ_k onto L; Gower and Hand [7] discuss interpolative biplots in the context of other forms of metric multidimensional scaling. Here, we are more concerned with predictive biplots that allow the values of variables to be associated with any point of L by reading off the appropriate markers in ξ_k. However, the whole process must be accomplished using only information provided in L. The nearness properties of ξ_k are the basis of prediction. The normal space N_μ at any marker μ on ξ_k contains all points that predict μ, including those points in L which belong to $L \cap N_\mu$. That is, N_μ contains all points that are nearer to the marker μ on ξ_k than to any other marker on ξ_k. The main problem of predictive biplots is to represent the intersection space by a single point in L, also marked with the label μ, in such a way that any point in L may be associated with the same marker in L as it would be by ξ_k in R_+^m. The kth biplot trajectory β_k is the locus in L of this point as μ varies.

In two dimensions, $(r = 2)$, with which we are mainly concerned, $L \cap N_\mu$. becomes a line:

$$\mathrm{L}(\mu): \quad l_1(\mu)z_1 + l_2(\mu)z_2 = m(\mu),$$

where the dependence of the coefficients on μ is emphasised. The actual functional forms of $l_1(\mu)$, $l_2(\mu)$ and $m(\mu)$ are given by (A.4) in the appendix. For simplicity we write the above as

$$\mathrm{L}(\mu): \quad l_1z_1 + l_2z_2 = m, \tag{5}$$

where l_1 and l_2 are direction cosines of the normal to $\mathrm{L}(\mu)$ in L satisfying

$$l_1^2 + l_2^2 = 1. \tag{6}$$

Thus, if we have a point (a_1, a_2) in L, the corresponding predicted value for the kth variable is given by the solution for μ of the nonlinear equation

$$l_1a_1 + l_2a_2 = m.$$

Gower [5] and Gower and Hand [7] show that a graphical solution to this equation may be provided by circular projection which is described next; Section 4 includes the first practical illustration of the method.

2.2.1. Circular projection

The projection of the origin, O, onto $\mathrm{L}(\mu)$ is the point (ml_1, ml_2) with which we associate the marker μ. Circular projection is defined by the locus of this point, as μ varies on ξ_k, giving a trajectory β_k for the kth variable. Suppose P is a point in L for which we wish to predict the value of its kth variable. β_k is used by drawing the circle on OP as diameter and reading off the marker at the point(s) where this circle intersects β_k; if there is more than one solution, take the one nearest to P. This is what is termed circular projection and it gives the

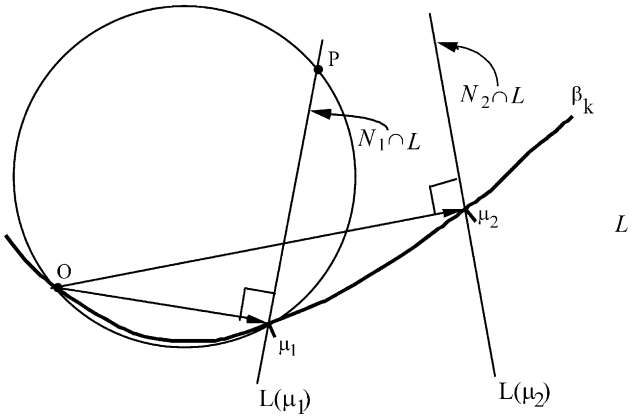

Fig. 1. Illustration of circular projection (modified from [7]).

correct prediction because diameters of circles subtend right angles on the circumference so ensuring that $L(\mu)$ is orthogonal to the line joining O to the marker for μ. The geometry of circular projection is shown in Fig. 1 for markers μ_1 and μ_2. The intersection of the circle with all p trajectories gives simultaneous predictions for all variables associated with P. When β_k is linear through the origin, as with classical biplots, circular projection is the same as orthogonal projection and circular projection then gives a convenient method for simultaneously obtaining and exhibiting all p orthogonal projections.

3. Normal projection

Predictive biplot trajectories based on the easily computed coordinates (ml_1, ml_2) of circular projection are not particularly convenient to use, unless with appropriate interactive computer facilities. Drawing circles freehand, or in the mind's eye, is unreliable and open to distortion. Classical linear predictive biplots are based on the simple notion of orthogonal projection from P onto linear biplot axes β_k $(k = 1, 2, \ldots, p)$ and it was shown above that prediction in R_+^m is also based on normal projection onto ξ_k $(k = 1, 2, \ldots, p)$. It is therefore natural to ask if predictive biplot trajectories based on normal projection might also be constructed in L. This requires the construction of a trajectory β_k that is normal to all the spaces $L \cap N_\mu$.

3.1. Normal projection in two dimensions

When $r = 2$, a trajectory normal to all the lines $L(\mu)$ of (5) is required. Suppose such a trajectory is defined by

$$\left.\begin{aligned} z_1 &= f_1(\mu) \\ z_2 &= f_2(\mu) \end{aligned}\right\} \tag{7}$$

and that it passes through the point $z_1 = \alpha_1, z_2 = \alpha_2$. This requires that

$$l_1(\mu)\alpha_1 + l_2(\mu)\alpha_2 = m(\mu)$$

which has root(s) $\mu = \mu_0$, say, and hence (α_1, α_2) is associated with the marker μ_0. We could start with any other point on $\mathrm{L}(\mu_0)$ and obtain a different trajectory. Although we present the results in general, for simplicity it will often be convenient to choose (α_1, α_2) to be the origin.

Writing $\dot{f}$ for $\frac{df}{d\mu}$, the tangent to the trajectory at μ has directions proportional to $(\dot{f}_1, \dot{f}_2)$, and the normality condition therefore requires

$$\frac{\dot{f}_1}{l_1} = \frac{\dot{f}_2}{l_2} = \lambda(\mu), \text{ (say)}. \tag{8}$$

Also (7) lies on $\mathrm{L}(\mu)$ and therefore

$$l_1 f_1 + l_2 f_2 = m. \tag{9}$$

For a diagram, the reader may refer ahead to that part of Fig. 2 that shows the normal projection trajectory labelled $\mathrm{N}(\mu)$. The point where this trajectory crosses $\mathrm{L}(\mu)$ has coordinates $(f_1(\mu), f_2(\mu))$ and the direction of the tangent at this point is given by $(\dot{f}_1(\mu), \dot{f}_2(\mu))$ which, by definition, is normal to $\mathrm{L}(\mu)$, as shown.

Differentiating (9) with respect to μ and substituting for f_2 from (9) and for $\dot{f}_2$ from (8) gives, after some rearrangement, the differential equation

$$\dot{f}_1 + \frac{l_1}{l_2}\left(\dot{l}_1 l_2 - \dot{l}_2 l_1\right) f_1 = \frac{l_1}{l_2}\left(\dot{m}\, l_2 - \dot{l}_2 m\right) \tag{10}$$

which on using (6) simplifies to

$$\dot{f}_1 - \frac{\dot{l}_2}{l_2} f_1 = \frac{l_1}{l_2}\left(\dot{m}\, l_2 - \dot{l}_2 m\right) \tag{11}$$

and hence

$$\frac{d}{d\mu}\left(\frac{f_1}{l_2}\right) = \frac{l_1}{l_2^2}\left(\dot{m}\, l_2 - \dot{l}_2 m\right). \tag{12}$$

Integrating (12) gives

$$\begin{aligned} f_1 &= l_2\left[\int_{\mu_0}^{\mu} \frac{l_1}{l_2^2}\left(\dot{m}\, l_2 - \dot{l}_2 m\right) d\mu + k\right] \\ &= l_2\left[\int_{\mu_0}^{\mu} l_1\, d\left(\frac{m}{l_2}\right) + k\right], \end{aligned} \tag{13}$$

where k is a constant of integration determined by the requirement that the trajectory passes through $f_1(\mu_0) = \alpha_1$, $f_2(\mu_0) = \alpha_2$, giving

$$k = \frac{\alpha_1}{l_2(\mu_0)}.$$

Thus finally,

$$\left.\begin{aligned} f_1 &= l_2\left[\int_{\mu_0}^{\mu} l_1\, d\left(\frac{m}{l_2}\right) + \frac{\alpha_1}{l_2(\mu_0)}\right] \\ f_2 &= l_1\left[\int_{\mu_0}^{\mu} l_2\, d\left(\frac{m}{l_1}\right) + \frac{\alpha_2}{l_1(\mu_0)}\right] \end{aligned}\right\} \tag{14}$$

gives the equation of a normal projection trajectory in two dimensions.

3.1.1. Linear case

When ξ_k is a linear coordinate axis, then $l_1(\mu)$ and $l_2(\mu)$ are constants and $m(\mu) = \kappa\mu$, where κ is a known factor of proportionality, chosen to give regularly spaced markers. Then (14) becomes

$$z_1 = f_1(\mu) = l_2 \int_{\mu_0}^{\mu} \kappa \frac{l_1}{l_2}\, d\mu + \alpha_1 = \kappa l_1(\mu - \mu_0) + \alpha_1$$

and

$$z_2 = f_2(\mu) = \kappa l_2(\mu - \mu_0) + \alpha_2.$$

Thus z_1 and z_2 are collinear on a line passing through (α_1, α_2) that is orthogonal to $l_1 z_1 + l_2 z_2 = \kappa\mu$ for all κ. This is as required for the classical linear biplot. μ_0 is given by $l_1\alpha_1 + l_2\alpha_2 = \kappa\mu_0$ and hence

$$\left.\begin{aligned} z_1(\mu) &= l_2^2\alpha_1 - l_1 l_2\alpha_2 + \kappa l_1\mu \\ z_2(\mu) &= l_1^2\alpha_2 - l_1 l_2\alpha_1 + \kappa l_2\mu \end{aligned}\right\} \tag{15}$$

gives the coordinates of the marker μ on the linear trajectory; usually, we set $\alpha_1 = \alpha_2 = 0$.

3.1.2. Circular case

Suppose the lines $\mathrm{L}(\mu)$ are radial with $l_1(\mu) = -\sin(\mu)$, $l_2(\mu) = \cos(\mu)$ and $m(\mu) = 0$. Thus, we have

$$\mathrm{L}(\mu) : \sin(\mu) z_1 - \cos(\mu) z_2 = 0.$$

We require a trajectory that is normal to all these radii and we shall suppose it passes through $\alpha_1 = 1$, $\alpha_2 = 0$ which corresponds to $\mu_0 = 0$. From (13), or directly from (8), we have that

$$\left.\begin{aligned} z_1(\mu) &= k\cos(\mu) \\ z_2(\mu) &= k\sin(\mu) \end{aligned}\right\} \tag{16}$$

and the boundary condition gives $k = 1$. Thus, the trajectory (16) is the parametric equation of a unit circle, as it obviously should be. The main interest of this example is to show that care has to be taken when either $l_1(\mu_0)$ or $l_2(\mu_0)$ is zero, in which case there are indeterminices in determining the constants of integration occurring in (14). One of the two integrals in (14) always has a determinate constant of integration so that $f_1(\mu)$, say, may be found and then (9) may be used to determine $f_2(\mu)$. Indeed, computationally, this approach is always to be preferred because it needs only one integration for every value of μ.

3.2. Normal projection in more than two dimensions

When $r = 3$, (8), (9) and (6) become

$$\frac{\dot{f}_i}{l_i} = \lambda(\mu) \quad \text{for } i = 1, 2, 3, \tag{17}$$

$$\sum_{i=1}^{3} l_i f_i = m \tag{18}$$

and

$$\sum_{i=1}^{3} l_i^2 = 1. \tag{19}$$

Differentiating (19) and using (17) gives

$$\sum_{i=1}^{3} \dot{l}_i l_i = \frac{1}{\lambda} \sum_{i=1}^{3} \dot{l}_i \dot{f}_i = 0. \tag{20}$$

The boundary condition that the trajectory passes through $(\alpha_1, \alpha_2, \alpha_3)$ gives

$$f_i(\mu_0) = \alpha_i \quad \text{for } i = 1, 2, 3,$$

where μ_0 is determined as a root of the equation

$$\sum_{i=1}^{3} l_i(\mu)\alpha_i = m(\mu). \tag{21}$$

With these preliminaries we proceed as in the two-dimensional case by differentiating (18) to give

$$\sum_{i=1}^{3} (\dot{l}_i f_i + l_i \dot{f}_i) = \dot{m} \tag{22}$$

which on substituting for $\dot{f}_2$ and $\dot{f}_3$ from (17) simplifies to

$$\sum_{i=1}^{3} \left(\dot{l}_i f_i + l_i \frac{l_i^2}{l_1} \dot{f}_1 \right) = \dot{m} \tag{23}$$

which on using (19) further simplifies to

$$\sum_{i=1}^{3}\left(\dot{l}_i f_i\right)+\frac{\dot{f}_1}{l_1}=\dot{m}. \tag{24}$$

Differentiating (24) and using (20) gives

$$\sum_{i=1}^{3}\left(\ddot{l}_i f_i\right)+\frac{l_1\ddot{f}_1-\dot{l}_1\dot{f}_1}{l_1^2}=\ddot{m}. \tag{25}$$

Eqs. (18), (24) and (25) may be put into the following matrix form:

$$\begin{pmatrix} l_1 & l_2 & l_3\\ \dot{l}_1 & \dot{l}_2 & \dot{l}_3\\ \ddot{l}_1 & \ddot{l}_2 & \ddot{l}_3\end{pmatrix}\begin{pmatrix} f_1\\ f_2\\ f_3\end{pmatrix}=\begin{pmatrix} m\\ \dot{m}-\frac{\dot{f}_1}{l_1}\\ \ddot{m}+\frac{\dot{l}_1\dot{f}_1}{l_1^2}-\frac{\ddot{f}_1}{l_1}\end{pmatrix}=\begin{pmatrix} m\\ \dot{m}-\lambda\\ \ddot{m}-\dot{\lambda}\end{pmatrix} \tag{26}$$

which may be abbreviated to

$$\mathbf{L}\mathbf{f}=\mathbf{g} \tag{27}$$

with solution

$$\mathbf{f}=\mathbf{L}^{-1}\mathbf{g}. \tag{28}$$

Clearly (28) defines three equations, the first of which is a nonhomogeneous second-order differential equation in f_1. The solution of this equation, subject to the boundary conditions (21), may be substituted into the second and third equations to obtain f_2 and f_3.

For general values of r we may proceed as above and arrive at (28) with

$$\mathbf{L}=\begin{pmatrix} l_1 & l_2 & l_3 & \cdots & l_r\\ \dot{l}_1 & \dot{l}_2 & \dot{l}_3 & \cdots & \dot{l}_r\\ \ddot{l}_i & \ddot{l}_2 & \ddot{l}_3 & \cdots & \ddot{l}_r\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ {}^{[r-1]}l_1 & {}^{[r-1]}l_2 & {}^{[r-1]}l_3 & \cdots & {}^{[r-1]}l_r\end{pmatrix}\quad\text{and}\quad \mathbf{g}=\begin{pmatrix} m\\ \dot{m}-g_1\\ \ddot{m}-g_2\\ \vdots\\ {}^{[r-1]}m-g_{r-1}\end{pmatrix}$$

with $g_1=\lambda$ and $g_2=\dot{\lambda}$ and where ${}^{[q]}l$ denotes differentiation q times. To determine g_q, differentiation of the qth equation yields

$$\sum_{i=1}^{r}{}^{[q]}l_i\, f_i={}^{[q]}m-\dot{g}_{q-1}-\sum_{i=1}^{r}\left({}^{[q-1]}l_i\, l_i\right)\frac{\dot{f}_1}{l_1}$$

so that

$$g_q=\dot{g}_{q-1}+\sum_{i=1}^{r}\left({}^{[q-1]}l_i\, l_i\right)\frac{\dot{f}_1}{l_1} \tag{29}$$

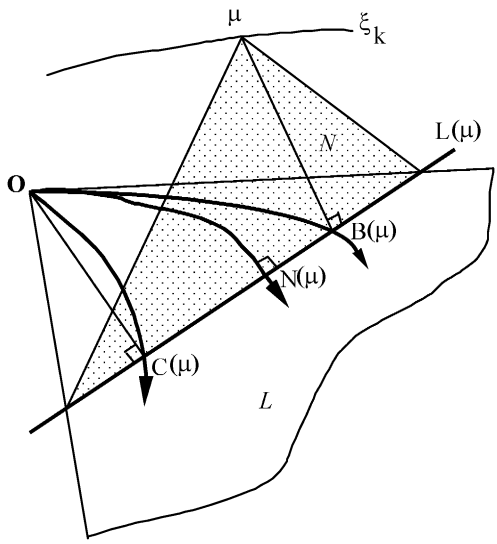

Fig. 2. The arrowed lines illustrate trajectories for three kinds of predictive biplot. B(μ) is the marker for μ at the nearest point (the back-projection of μ) on L(μ) to μ on ξ_k. C(μ) is the marker for μ to be used for circular projection. N(μ) is the marker for μ to be used for normal projection. The loci of these points with varying μ defines the three trajectories.

remains a function of $\frac{\dot{f}_1}{l_1}$ for all values of q. Thus, for any value of r, f_1 may be found as a solution to a differential equation but the computational cost is high. In practice, we are nearly always concerned with the two-dimensional case, which requires only the integrals (14). While more complicated than the simple equations for circular projection, the computation of trajectories for two-dimensional normal-projection is perfectly feasible (Fig. 2).

3.3. Comparison of predictive biplots

Fig. 2 shows three types of predictive biplot axes: trajectories for circular projection, normal projection and back projection. The origin is shown at O corresponding to a fixed point (α_1, α_2) and a value μ_0 of the kth variable. For circular projection, a simple generalisation of the argument of Section 2.1 shows that the coordinates of C(μ) are

$$\mathrm{C}(\mu) : (m - \mathbf{l}'\boldsymbol{\alpha})\mathbf{l}, \tag{30}$$

where $\boldsymbol{\alpha}'$ is the r-dimensional fixed point $(\alpha_1, \alpha_2, \ldots, \alpha_r)$ and $\mathbf{l}' = (l_1, l_2, \ldots, l_r)$.

The coordinates of N(μ) are the concern of Section 4 and are given by the solution to (14) when $r = 2$ and, in general, by (28).

The back projection B(μ) is the point on L(μ) that is nearest to the marker μ on ξ_k. This is the same as the projection onto L(μ) of the projection $\mathbf{z}_\mu$ of $\mathbf{y}_\mu$ (given by (4)) onto L and

has coordinates

$$\mathrm{B}(\mu) : m\mathbf{l} + (\mathbf{I}_r - \mathbf{l}\mathbf{l}')\mathbf{z}_\mu. \tag{31}$$

The back-projection of a point on ξ_k is unique so, unlike the other two trajectories $\mathrm{B}(\mu)$ cannot be made to pass through any point $\boldsymbol{\alpha}$ in L. That it is shown in Fig. 2 as passing through O, implies that O has been chosen as the back-projection of some marker on ξ_k, such as the zero value of the kth variable.

In the linear case and when α is chosen as the origin (i.e. the centroid of $\mathbf{Y}$), $\mathrm{C}(\mu)$, $\mathrm{N}(\mu)$ and $\mathrm{B}(\mu)$ coincide. $\mathrm{B}(\mu)$ is easy to compute and in view of its nearness properties, should be important. It may be regarded as giving the best representation attainable in L of the kth coordinate trajectory ξ_k. Unfortunately, it does not seem possible to use the locus of $\mathrm{B}(\mu)$ to reconstruct $\mathrm{L}(\mu)$ as we have seen is possible for $\mathrm{C}(\mu)$ and $\mathrm{N}(\mu)$. The reason for this is that the constructions of $\mathrm{C}(\mu)$ and $\mathrm{N}(\mu)$ use angles defined within L whereas $\mathrm{B}(\mu)$ calls on information available only in R^m_+. In the following example, we explore its possible use as an approximation to the trajectory for normal projection.

4. Example

Table 1 shows the values of four variables for 21 types of aircraft [13]. Inspection shows that the variable PLF is not commensurable with the others so, just as with a principal components analysis, some form of normalization is necessary before calculating a distance suitable for multidimensional scaling. We have chosen an unusual form of normalization, due to Clark [1], which has some popularity in studies in plant ecology and which defines ddistance by setting in (1):

$$d_k(x_{ik}, x_{jk}) = \left(\frac{x_{ik} - x_{jk}}{x_{ik} + x_{jk}}\right)^2$$

giving

$$\mathbf{d}_k = \left\{\frac{x_{ik} - \mu}{x_{ik} + \mu}\right\}^2 \quad \text{and} \quad \dot{\mathbf{d}}_k = -\left\{4x_{ik}\frac{(x_{ik} - \mu)}{(x_{ik} + \mu)^3}\right\}. \tag{32}$$

Gower and Legendre [9] showed that this distance is Euclidean embeddable (assumption ii) when, as in Table 1, the values of $\mathbf{X}$ are all nonnegative. It is a distance that gives more weight to differences between two small-sized sample-units than to the same difference between two large-sized sample-units. Standardising variables has no effect but translating the origin does and therefore with interval-scales (32) should be used with caution. Note that when $\mu = 0$ then $d_{ik} = 1$ irrespective of the value of x_{ik}, so some unusual behaviour can be expected in the vicinity of zero on any variable—see the variable PLF for aircraft labelled g and r in Table 1.

The coordinates $\mathbf{Z}$ in $r = 2$ dimensions were determined by principal coordinates analysis so are given by the first two columns of $\mathbf{Y}$. To determine a point on the kth trajectory requires the values of l_1, l_2 and m given by (A.4) of the appendix, remembering to impose

Table 1
Information on 21 aircraft in chronological order of first flight

	Aircraft	SPR	RGF	PLF	SLF
a	FH-1	1.468	3.30	0.166	0.10
b	FJ-1	1.605	3.64	0.154	0.10
c	F-86A	2.168	4.87	0.177	2.90
d	F9F-2	2.054	4.72	0.275	1.10
e	F-94A	2.467	4.11	0.298	1.00
f	F3D-1	1.294	3.75	0.150	0.90
g	F-89A	2.183	3.97	0.000	2.40
h	XF10F-1	2.426	4.65	0.117	1.80
i	F9F-6	2.607	3.84	0.155	2.30
j	F100-A	4.567	4.92	0.138	3.20
k	F4D-1	4.588	3.82	0.249	3.50
m	F11F-1	3.618	4.32	0.143	2.80
n	F-101A	5.855	4.53	0.172	2.50
p	F3H-2	2.898	4.48	0.178	3.00
q	F102-A	3.880	5.39	0.101	3.00
r	F-8A	0.455	4.99	0.008	2.64
s	F-104A	8.088	4.50	0.251	2.70
t	F-105B	6.502	5.20	0.366	2.90
u	YF-107A	6.081	5.65	0.106	2.90
v	F-106A	7.105	5.40	0.089	3.20
w	F-4B	8.548	4.20	0.222	2.90

SPR—Specific power (proportional to power per unit weight), RGF—flight range factor, PLF—payload as a fraction of gross weight of aircraft, SLF—sustained loan factor.

the normalization (6). Apart from this normalzation, (A.4) gives

$$l_j = -\lambda_j^{-1}\mathbf{y}_j'\,\dot{\mathbf{d}}_k = 4\lambda_j^{-1}\sum_{i=1}^{n} y_{ij}x_{ik}\frac{(x_{ik}-\mu)}{(x_{ik}+\mu)^3} \quad \text{and}$$

$$m = \frac{1}{n}\mathbf{1}'\,\dot{\mathbf{d}}_k = -\frac{4}{n}\sum_{i=1}^{n} x_{ik}\frac{(x_{ik}-\mu)}{(x_{ik}+\mu)^3},$$

where $\mathbf{y}_j$ is the jth column of $\mathbf{Y}$; in this example we are concerned only with $j = 1, 2$. Thus, for the kth variable, l_j is determined solely from the jth dimension of the ordination and the values of the kth variable. The algebraic prediction of the value μ of the kth variable associated with the point with coordinates (a_1, a_2) requires the solution of the nonlinear equation

$$\mathbf{w}'\,\dot{\mathbf{d}}_k = 0,$$

where the column-vector $\mathbf{w}$ has elements $\left\{\frac{a_1}{\lambda_1}y_{1i} + \frac{a_2}{\lambda_2}y_{2i} + \frac{1}{n}\right\}$. Circular and normal projection provide graphical tools for solving this equation.

For circular projection, the trajectories through the origin are the loci of the point $f_1 = ml_1$, $f_2 = ml_2$ $(k = 1, 2, 3, 4)$. These are shown in Fig. 3 for a series of values of μ together

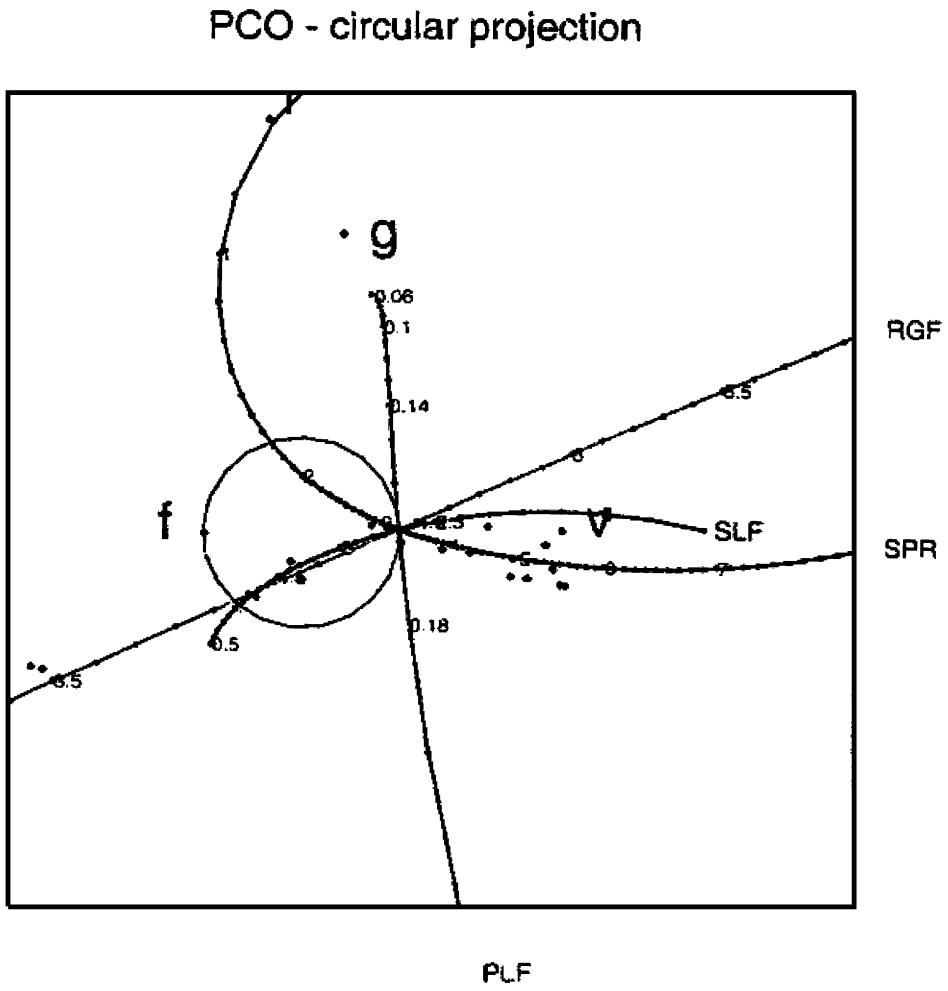

Fig. 3. Two-dimensional principal coordinates analysis for the data of Table 1 with distance given by (32). The calibrated trajectories for circular projection are shown together with the projection circle for aircraft *f*.

with the positions of the samples as given by the two-dimensional principal coordinates analysis which accounts for 62.7% of the total ddistance.

The effect of the choice of distance is striking. The aircraft of Table 3 are listed in chronological order so most values of the variables increase with time. In general. the aircraft get more powerful (SPR), can travel further (RGF), have increased payload (PLF), which they can carry for longer times (SLF). The result of using (32) is that distances between pairs of the earlier aircraft tend to be greater than the distances between pairs of the later aircraft—which bunch on the right-hand side of the figure. Aircraft *g* and *r* are anomalous, but the others show the expected trend with time. Both *g* and *r* have very low

Table 2
The positive roots of $\mathbf{1}'\mathbf{d}_k = 0$

Variable k	SPR	RGF	PLF	SLF
Root μ_0	3.24	4.44	0.168	2.34
Mean	3.83	4.49	0.167	2.28

values of PLF while r is also low on SPR and, as stated above, the effect of a zero (low) value in a variable is a unit (close to unit) contribution to distance from all other aircraft. Thus, because a unit value is maximal, aircraft with low values can be expected to be peripheral in the overall display. As $\mu \to \infty$ we also have that $d_k(x_{ik}, \mu) \to 1$, which we conjecture is conducive to the formation of horseshoe effects for the higher values of scales but as we do not extrapolate beyond the range of data-values we do not encounter this phenomenon. The discussion of the trajectories in Fig. 3 is deferred until we have the corresponding figure for normal projection.

For simplicity, we have chosen that the trajectories for normal prediction should also pass through the origin of the ordination, so making the constants of integration in (14) zero. A line through the origin has $m = 0$ so, from (A.4) of the appendix, μ_0 is determined for the kth variable as the solution of the nonlinear equation $\mathbf{1}'\dot{\mathbf{d}}_k = 0$. It can be shown that this equation has only one positive root and it lies in the range $x_{1k} < \mu_0 < x_{nk}$. The root is easily found by the bisection method and the values for the four variables are shown in Table 2.

From Table 2 we see that μ_0 is near the mean for each variable; in the linear case the two coincide. The functional forms of l_1, l_2 and m remain as for circular projection but for normal projection must be substituted into (14) to determine f_1 and f_2. However, only one integral need be evaluated, say the first, because

$$l_1 f_1 + l_2 f_2 = m(\mu)$$

allows f_2 to be derived when f_1 is known. The resulting trajectories for normal projection are shown in Fig. 4.

We discuss the trajectories of Figs. 3 and 4 simultaneously. First, recall that the predictions derived from both diagrams must be the same because they give alternative methods for constructing the same lines $\mathrm{L}(\mu)$. The trajectories for circular and normal projection have much in common but those for normal projection tend to be less curved at the ends of their ranges. In both cases the variable RGF turns out to have a trajectory very close to linearity and with little departure from regular calibration. In both cases, the trajectory for PLF is close to being linear but its calibrations are irregular and become crowded as zero is approached. This is the result of the special effect of zero values in (32). Indeed, with PLF, Table 1 shows a zero reading for aircraft f so the vector $\mathbf{d}_k$ is not defined at $\mu = 0$ and the calculation of the trajectory becomes unstable as zero is approached; a similar crowding of calibrations occurs for low values of the variable SLF. The trajectories for SLF and SPR are highly curvilinear.

Fig. 3 shows the circular predictions associated with aircraft f and Fig. 4 the corresponding normal predictions. The algebraically calculated predictions for f and v are given in

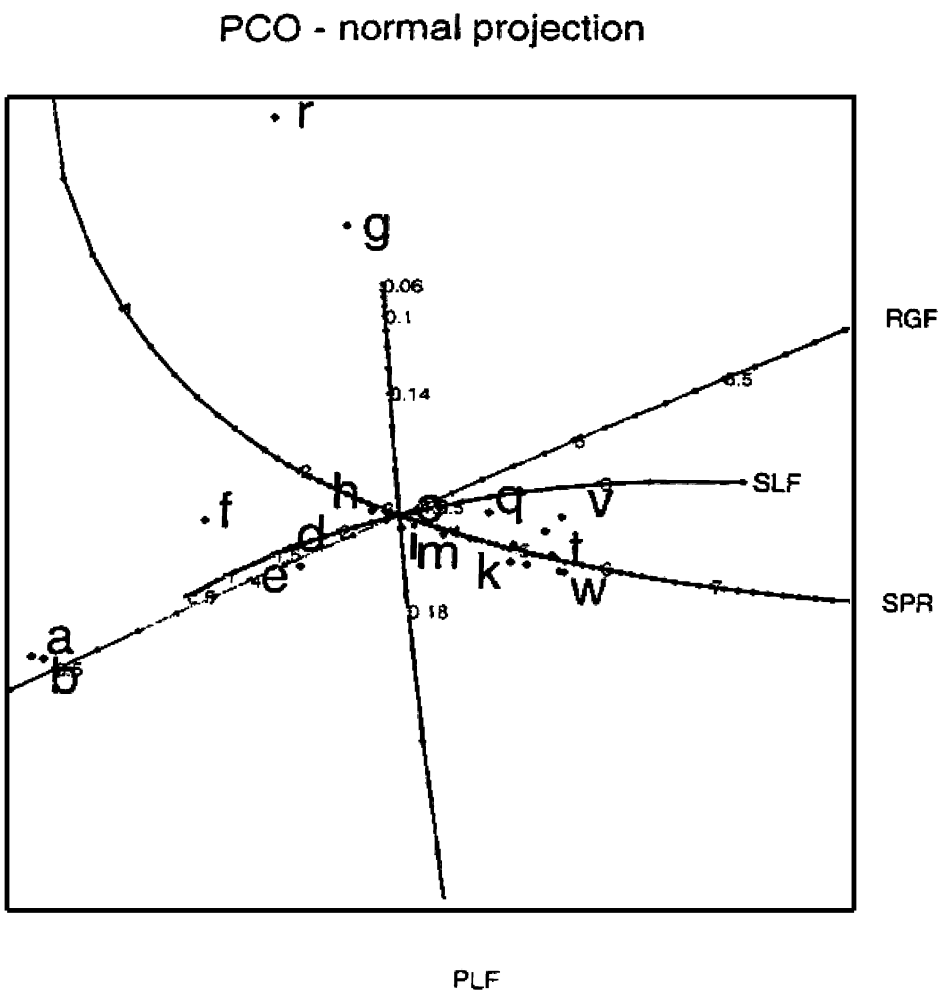

Fig. 4. Two-dimensional principal coordinates analysis for the data of Table 1 with distance given by (32). The calibrated trajectories for normal projection are shown together with the normal projections for aircraft *f*.

Table 3 and it can be seen that the graphical results agree with these values as well as can be expected, given the restricted accuracy of visual inspection.

It can be seen that the predicted values have the right orders of magnitude, and have an accuracy consistent with the 66.7% fit of the two-dimensional representation. Other choices of distance would do better. For example, the principal components analysis given by Gower and Hand [7], equivalent to the Eckart–Young [2] approximation, gives much better predictions in the least-squares sense. The essential difference is that with the current analysis it is the distance (32) that has priority in determining the approximation. As we have seen, this choice of distance effectively downweights large values of a variable relative

Table 3
Observed and algebraically calculated predicted values for aircraft f and v

Variable	Aircraft			
	f		v	
	Observed	Predicted	Observed	Predicted
SPR	1.294	1.66	7.105	5.42
RGF	3.75	3.95	5.40	4.87
PLF	0.150	0.166	0.089	0.18
SLF	0.90	1.07	3.20	2.89

to low values and if that is deemed desirable then the Eckart–Young theorem, at least in its usual unweighted form, is no longer relevant. Indeed, we may view our predictions as one of a family of matrix approximations, starting with the Eckart–Young approximation which gives the rank r approximation $\hat{\mathbf{X}}_W$ which minimises $trace(\mathbf{X}-\hat{\mathbf{X}})'(\mathbf{X}-\hat{\mathbf{X}})$ with the generalization which minimises $trace(\mathbf{X}-\hat{\mathbf{X}})'\mathbf{W}^{-1}(\mathbf{X}-\hat{\mathbf{X}})$ "in the metric $\mathbf{W}$", where $\mathbf{W}$ is positive definite (see e.g. [7]) and in statistical applications is typically a within-groups dispersion matrix, with the further generalization which gives the approximation $\hat{\mathbf{X}}_D$ described above. $\hat{\mathbf{X}}_D$ may be said to minimise in the Euclidean embeddable distance $\mathbf{D}$. Note that although our representations are r-dimensional, generally $rank(\hat{\mathbf{X}}_D) > r$.

Fig. 5 shows the back-projection trajectories for the same data, as in Figs. 3 and 4. As explained above, these trajectories are the points in L that are closest to the trajectories in R_+^m and are therefore unique and cannot be made to be concurrent. It may be verified that the same markers for a variable in Figs. 3–5 are collinear—on the line $\mathrm{L}(\mu)$ as shown in Fig. 2. With this example, the trajectories for back-projection are more similar to those for circular projection than for normal projection but no general conclusion can be made. With orthogonal axes, we know that a point lying on one axes implies that its values for all other variables must be close to zero, or whatever other values are associated with the origin. With the closeness property of back-projection it might be expected that a point lying on or close to a trajectory would convey similar useful information This is not so, even for the exact trajectories ξ_k where, because of the non-orthogonality of the trajectories, projections onto the other trajectories may be far from the origin. Not only does this observation apply to all the nonlinear trajectories discussed here but also to linear nonorthogonal classical biplots. All that can be said is that if a point lies on a back-projection trajectory then it is closer to the marker μ on ξ_k than is any other point on $\mathrm{L}(\mu)$.

5. Conclusion

We have presented a general methodology for investigating the consequences of using nonlinear functions of basic variables to provide distances that are then approximated by multidimensional scaling. The calibrated biplot allows the effects of defining different distances both by assessing the extent of nonlinearity in its trajectory and by the evenness in

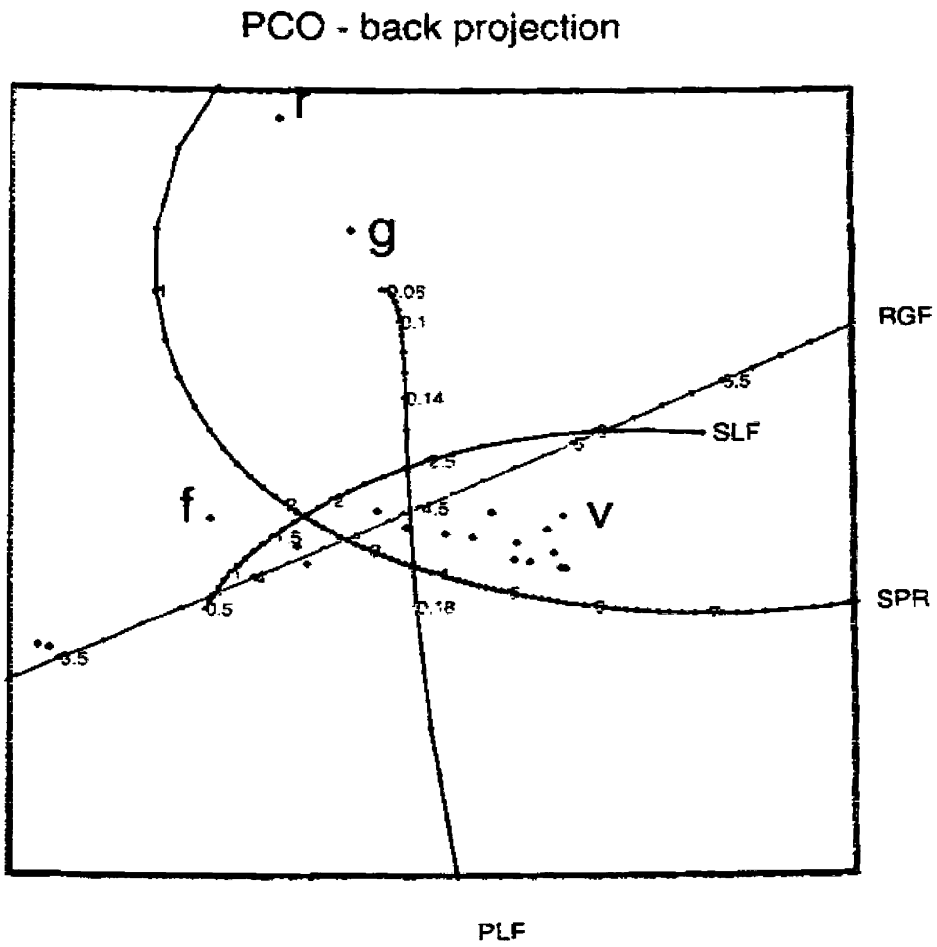

Fig. 5. Two-dimensional principal coordinates analysis for the data of Table 1 with distance given by (32). The calibrated trajectories for back projection are shown.

the distribution of its markers. Sometimes, properties may be found that could be deemed to outweigh other desirable properties of the chosen distance. In our example with the ddistance (32) we saw that zero and small values of a variable had a marked effect on the multidimensional scaling. Under some circumstances this effect might be regarded as a good way of drawing attention to anomalous low values but under other circumstances it might be regarded as indicating that the chosen distance function should be rejected. This example exhibited the highly nonlinear trajectories associated with some of the variables, thus warning against the indiscriminant use of the regression method (see e.g. [11] or [7]) for providing linear biplot axes for any nonlinear method of multidimensional scaling. Even an irregularly calibrated linear axis such as PLF is inconsistent with use of the regression method.

We could give algebraic results because of the restrictions imposed by the assumptions made in Section 2.1, especially that of Euclidean embeddability which underpins the principal coordinate/classical scaling approach to MDS. There are, of course, many other methods both of metric and nonmetric MDS. Just as the configurations found by the different methods tend to be fairly similar, we would expect similar effects on nonlinearity to those we have found. With some types of MDS, parts of our methodology may continue to be used. Thus, when the exact representation of $\mathbf{Y}$ and its associated trajectories in R_+^m exist, we may seek to embed any MDS configuration $\mathbf{Z}$, however found. One way of doing this is to find the projection matrix $\mathbf{P}$ that minimises $\|\mathbf{YP} - \mathbf{Z}\|^2$. We may then identify L with the space spanned by the columns of $\mathbf{YP}$ and proceed as in Sections 2 and 3. Another possibility is when $\mathbf{Y}$ is obtained by optimal nonlinear transformations of the columns of $\mathbf{X}$, chosen to optimise an MDS criterion that matches the distances between the rows of $\mathbf{Y}$ with corresponding distances between the rows of $\mathbf{Z}$. When the distances between the rows of $\mathbf{Y}$ are chosen to be Pythagorean in R^p relative to orthogonal axes, then $\mathbf{Z}$ may be embedded by projection, as described above, giving linear biplot axes in L. However, the nonlinear transformations from $\mathbf{X}$ to $\mathbf{Y}$ induce irregular calibrations on these linear axes (see e.g. [10]). If non-Pythagorean distances are chosen, we arrive at nonlinear trajectories whose irregular calibrations have to be labelled to accommodate the further nonlinear effect of the transformation from $\mathbf{X}$ to $\mathbf{Y}$. When the distance function is not Euclidean, in principal we can always use the pseudo-sample approach to provide calibrated nonlinear trajectories in L. However, the use of concepts like, nearness, projection and vector-sum then are hard to justify and with no proper linkage between the sample points and the trajectories that represent the variables, interpretability is lost.

We have shown how to compute biplot trajectories for prediction by normal projection. These trajectories are easy to use visually and are an obvious generalization of the familiar usage with linear biplots. Circular projection may be viewed as a generalization of the linear usage. On a computer, circular and normal projection are equally convenient to use but circular projection is the more efficient computationally. With hard copy, predictions given by normal projection are the more convenient.

We have noted the relationship of this work with generalisations of the least-squares rank r approximation of the matrix $\mathbf{X}$.

AppendixA. Equation of the normal plane N_μ

To simplify notation, we take $m = n - 1$. If $m < n - 1$, terms like y_n should be replaced by y_{m+1} but there is no substantive effect.

We shall write the normal plane N_μ as $\mathbf{t}'\mathbf{y} + t_n y_n = \mathbf{t}'\mathbf{y}_\mu + t_n y_n(\mu)$ where $\mathbf{y}' = (y_1, y_2, \ldots, y_{n-1})$ are variables in the first $n-1$ dimensions, y_n refers to the nth dimension in R_+^m and $\mathbf{y}_\mu$ are the coordinates at μ given by (4). The coefficients $\mathbf{t}' = (t_1, t_2, \ldots, t_{n-1})$ are in the directions of the tangent to ξ_k at $\mathbf{y}_\mu$ and hence are obtained by differentiating the first part of (4) to give

$$\mathbf{t} = \dot{\mathbf{y}}_\mu = -\frac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{Y}'\dot{\mathbf{d}}_k \tag{A.1}$$

with the direction $t_n = \dot{y}_n(\mu)$ in the residual dimension in R^m_+ obtained from the second part of (4) as

$$2y_n(\mu)t_n = 2y_n(\mu)\,\dot{y}_n(\mu) = \frac{1}{n}\mathbf{1}'\,\dot{\mathbf{d}}_k - 2\mathbf{y}'_\mu\,\dot{\mathbf{y}}_\mu.$$

Multiplying $\mathbf{y}_\mu$ given by (4) by $\dot{\mathbf{y}}_\mu$ given by (A.1) routinely yields

$$\mathbf{y}'_\mu\,\dot{\mathbf{y}}_\mu = \frac{1}{4}\left[\sum_{h=1}^{p}\mathbf{c}_h - \mathbf{c}_k + \mathbf{d}_k - \frac{2}{n}\mathbf{D1}\right]'\mathbf{B}^-\,\dot{\mathbf{d}}_k$$

and therefore

$$2y_n(\mu)t_n = \left(\frac{1}{n}\mathbf{1}' - \frac{1}{2}\left[\sum_{h=1}^{p}\mathbf{c}_h - \mathbf{c}_k + \mathbf{d}_k - \frac{2}{n}\mathbf{D1}\right]'\mathbf{B}^-\right)\dot{\mathbf{d}}_k \tag{A.2}$$

which using the second part of (4) allows $\mathbf{t}_n$ to be calculated. In establishing (A.2) we have used the result that the Moore–Penrose inverse $\mathbf{B}^-$ is given by $\mathbf{B}^- = \mathbf{Y}(\mathbf{\Lambda}^-)^2\mathbf{Y}'$ (see [7, p. 248]).

Eqs. (A.1) and (A.2) give the directions of the tangent and hence the plane N_μ normal to the trajectory at μ. From the second part of (4), we have the constant term

$$\mathbf{t}'\mathbf{y}_\mu + t_n y_n(\mu) = \frac{1}{2}\frac{d(y_n^2(\mu) + \mathbf{y}'_\mu\mathbf{y}_\mu)}{d\mu} = \frac{1}{2n}\mathbf{1}'\,\dot{\mathbf{d}}_k. \tag{A.3}$$

Thus N_μ may be written

$$-\mathbf{y}'\mathbf{\Lambda}^{-1}\mathbf{Y}'\,\dot{\mathbf{d}}_k + 2t_n y_n = \frac{1}{n}\mathbf{1}'\,\dot{\mathbf{d}}_k. \tag{A.4}$$

When L is already embedded optimally in R^m_+, the intersection $\mathrm{L}(\mu)$ with L, is given by taking the first r terms $(y_1, y_2, \ldots, y_r)$ of (A.4). Thus (A.4) gives the coefficients $l_i(\mu)$ and $m(\mu)$ apart from a normalizer required to convert to direction cosines. Note that $t_n y_n$ does not enter into the intersection, because y_n occurs in the dimension of R^m_+ that is orthogonal to R^m and hence is also orthogonal to its subspace L.

When L has to be embedded in R^m_+ by the projection $\mathbf{P}$ that minimizes $\|\mathbf{YP} - \mathbf{Z}\|$, as outlined in Section 5, we need to ensure that (A.4) is expressed in terms of the coordinate system of $\mathbf{Z}$ in L. This requires a rotation of the axes of R^m, but because the intersection requires only the first r dimensions after rotation, we find that $\mathbf{y}'\mathbf{P} = \mathbf{z}$, say, gives the r-dimensional coordinates and the linear coefficients $\mathbf{t}$ are replaced by $\mathbf{P}'\mathbf{t}$. With this small change, we may proceed as before.

References

[1] P.J. Clark, An extension of the coefficient of divergence for use with multiple characters, Copeia 2 (1952) 61–64.

[2] C. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika 1 (1936) 211–218.

[3] K.R. Gabriel, The biplot graphical display of matrices with applications to principal components analysis, Biometrika 58 (1971) 453–467.
[4] J.C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (1964) 325–338.
[5] J.C. Gower, Recent advances in biplot methodology, in: C.M. Cuadras, C.R. Rao (Eds.), Multivariate Analysis: Future Directions, Vol. 2, Elsevier Science Publishers, Amsterdam, 1993, pp. 295–325.
[6] J.C. Gower, A general theory of biplots, in: W.J. Krzanowski (Ed.), Recent Advances in Descriptive Multivariate Analysis, Oxford University Press, Oxford, 1995, pp. 283–303.
[7] J.C. Gower, D.J. Hand, Biplots, Chapman & Hall, London, 1996.
[8] J.C. Gower, S.A. Harding, Nonlinear biplots, Biometrika 75 (1988) 445–455.
[9] J.C. Gower, P. Legendre, Metric and Euclidean properties of dissimilarity coefficients, J. Classification 3 (1986) 5–48.
[10] J.C. Gower, J.J. Meulman, G.M. Arnold, Non-metric linear biplots, J. Classification 16 (1999) 181–196.
[11] W.J. Kruskal, M. Wish, Multidimensional scaling, Sage University Papers on Quantitative Applications in the Social Sciences, Number 07-011, Sage Publications, Newbury Park, CA, 1978.
[12] I.J. Schoenberg, Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espaces vectoriels distanciés applicables vectoriellement sur l'espace de Hilbert", Ann. Math. 36 (1935) 724–732.
[13] W. Stanley, M. Miller, Measuring technological change in jet fighter aircraft, Report No. R-2249-AF, Rand Corporation, Santa Monica, CA, 1979.
[14] W.S. Torgerson, Theory and Methods of Scaling, Wiley, New York, 1955.